The work reported in this document was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology. This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract F19628-76-C-0002 (ARPA Order 2006).

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Raymond L. Loiselle

Raymond L. Loiselle, Lt. Col., USAF
Chief, ESD Lincoln Laboratory Project Office

Technical Note

1977-5

S. Seneff

Real Time Harmonic Pitch Detector

26 January 1977

Prepared for the Defense Advanced Research Projects Agency
under Electronic Systems Division Contract F19628-76-C-0002 by

Lincoln Laboratory

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LEXINGTON, MASSACHUSETTS

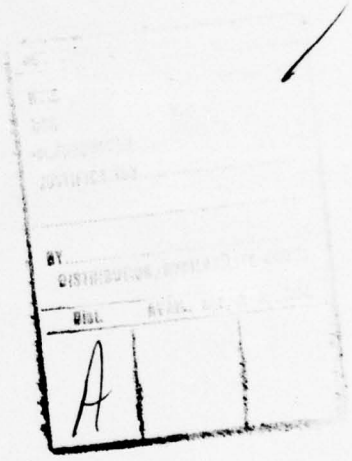
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LINCOLN LABORATORY

REAL TIME HARMONIC PITCH DETECTOR

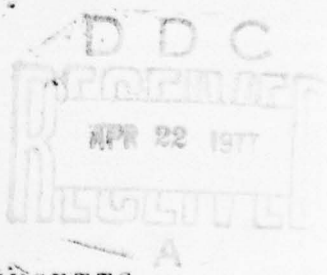
S. SENEFF

Group 24



TECHNICAL NOTE 1977-5

26 JANUARY 1977



LEXINGTON MASSACHUSETTS

ABSTRACT

A real time harmonic pitch detection algorithm has been developed on the Lincoln Digital Voice Terminal (LDVT). The algorithm was designed to be fast and to perform well when the input speech is degraded (i.e., telephone quality) or corrupted with acoustically coupled noise. The algorithm determines the fundamental frequency from the spacing between harmonics in a selected portion of the spectrum. The algorithm was incorporated into a real time linear prediction vocoder and compared favorably in informal listening tests with the Gold-Rabiner time domain detector under a variety of adverse conditions.

CONTENTS

I. INTRODUCTION

The speech waveform can be modelled as the response of the vocal tract filter to a source which is a periodic sequence of pulses during voiced segments or a random noise during unvoiced segments. The periodic pulses occur as a consequence of the opening and closing of the glottis, and the frequency of the periodicity is often referred to as the pitch.* The noise source is a consequence of a narrow constriction at some point in the vocal tract. The model is a simplification, for certain sounds, such as /v/, are driven by both a periodic and a noise source simultaneously. However, the model has proved to be sufficiently accurate that acceptable vocoders can be constructed based on the simple concept of either a periodic or noise source, but not both, at a given instance of time.

Once the model has been accepted, the difficult task is to determine the periodicity of the source when the speech

---

* Pitch is more strictly defined as the perceived rather than the generated frequency; the latter has been given various names, such as, laryngeal frequency, voice fundamental frequency, etc., but all are rather awkward compared to "pitch".

is voiced, and to determine that there is no periodicity when it is unvoiced. It is a task to which considerable attention has been devoted in the past, and as a consequence there are many published papers available on the subject of pitch extraction<2-7,9,10,12>.

All pitch detectors can be placed in one of two categories, time domain and frequency domain. Time domain pitch detectors deal directly with the speech waveform, and as such are relatively fast, since very little preprocessing of the signal is required. Most frequency domain detectors require an abundance of time and memory storage to obtain the spectral information over a sufficiently long time window and with adequate spectral resolution. These methods are therefore often not realizable in real-time vocoder implementations or only realizable at the cost of excessive quantization of the pitch.

Pitch detection is generally good when the input signal is intact and noise-free. However, distortions, filters and noise tend to obscure the pitch information and cause most pitch detectors to break down, sometimes severely. Since in the real world the signal is often corrupted, it was felt that an algorithm designed to be robust against degradations would be a significant new contribution.

We were particularly interested in coping with degradations caused by a) passage of the speech through the public telephone system prior to pitch detection and b) acoustically coupled noise backgrounds. From a previous effort <11>, we had the capability to simulate in real time the filtering, phase distortion, phase jitter, and non-linear distortion effects of a telephone system. In addition, we had available test material wherein the noise background of a large jet airplane was incorporated into the recording. Unfortunately, we did not have time to assess the effects of the carbon button microphone of conventional telephone terminals.

The algorithm thus developed is a frequency domain technique which however restricts itself to a selected portion of the frequency band below 1100 Hz. Digital signal processing tricks were used to obtain the desired spectral region with minimal computation time. Pitch is determined from spacing between peaks in this region, using an iterative method. The buzz-hiss decision makes use of none of the standard indicators such as energy ratios and zero crossing density, as these parameters are highly susceptible to noise and distortion. Instead, continuity of the pitch track is the only parameter used to determine voicing, other than a very conservative silence threshold. The algorithm has been incorporated into a real time linear prediction

vocoder implemented on the Lincoln Digital Voice Terminal (LDVT) <1>.

II. PREPROCESSING

In order to obtain an accurate pitch estimate from the spectral information it is necessary to begin with a spectrum with good frequency resolution, but one which spans a sufficient block of the frequency space for there to be at least two harmonics present over the range available. Since a pitch value of 350 Hz is not unreasonable for a female, the spectral region to be analyzed must be at least 700 Hz wide. Within this range one can arbitrarily choose an FFT size to yield the desired frequency spacing between samples at the expense of computer time. It was decided to compute a 128 point FFT to yield a spectrum spanning 840 Hz with a resulting frequency spacing between samples of 6.6 Hz, which appears to be adequate resolution for our purposes.

Since the pitch information must be extracted from precisely the 840 Hz region chosen, it is expedient to carefully select that region which is most likely to yield robust harmonics. Since the telephone filter removes the signal below about 300 Hz, one need not waste space on the

low end of the frequency spectrum. However, as one advances to increasingly higher frequencies, the spectrum becomes more and more ragged, and the harmonics increasingly difficult to extract.

The region selected was 210 to 1050 Hz. These particular numbers were arrived at in large part because of appropriate tricks that could be used to extract precisely this piece of the spectrum. The original speech waveform was analog filtered and sampled at 130 usec intervals, yielding a signal containing frequencies up to 3780 Hz, which could be used as input both to the linear prediction analysis and to the pitch extractor.

The first step in the pitch extraction is to filter the speech down to 1260 Hz and downsample, throwing away two out of every three samples(Figure 1). For this purpose a finite impulse response filter seemed to be a good choice. Since FIR filters have only zeros, one need compute outputs only at the downsampled rate, which in our case represents a three to one savings in time. Furthermore, FIR filters are implementable using charge coupled devices (CCD's), a potentially fast and inexpensive computational source.

We now have a waveform, s1(n), which contains information from -1260 to +1260Hz. We know that, since the waveform is real, the negative frequency information is

redundant. Digital signal processing theory tells us that if we multiply each sample of the waveform s1(n) by $e^{j\omega n}$, we will cause the spectrum to be rotated by ω in the z-plane. By choosing $\omega=90°$, we cause the spectrum to be rotated such that 1260/2=630 Hz is at the origin (Figure 2). Now a second pass of this complex s2(n) through the same filter, with 3 to 1 downsampling again, will yield a complex waveform, s3(n), containing frequencies up to 1260/3=420 Hz. However, because of the rotated spectrum, 630 Hz in the original waveform corresponds to zero Hz in s2(n), and thus our doubly downsampled complex waveform contains the information from (630-420) Hz to (630+420) Hz in the original speech, which is the desired spectral region.

Choosing $\omega=90°$ has certain advantages in terms of speed. Multiplication by $e^{j\omega n}$ involves only data transfer rather than complex multiplies, since the sine and cosine of multiples of 90° are always either ± 1 or 0. Furthermore, as a consequence, each sample of s2(n) is either purely real or purely imaginary. One can therefore use simple tricks in the implementation of the FIR filter so that filtering of this complex waveform takes essentially no more time than would filtering a real waveform.

Pitch detection generally requires a long time window of speech in order to assure at least two periods of a low

pitched voice. Fortunately, the doubly downsampled signal consists of samples which are spaced by 132x3x3 usec, or 1.188 msec. Only 32 samples of this waveform are required to yield 38 msec of data, a time window that is sufficient to encompass two periods for pitches of up to 19 msec, or 53 Hz, a very deep male voice.

The 32 most recent samples of s3(n) are windowed using a standard Hanning window and then filled out with zeros to make a 128 point input buffer for the FFT. Because one fourth of the input samples are zero, the FFT computation time can be reduced by essentially skipping the first 2 stages. The resulting spectrum contains the information in the original speech signal from 210 to 1050 Hz, as desired, and is ready, after the computation of the magnitude spectrum from real and imaginary components, to be processed for harmonic detection.

III. PEAK PICKING ALGORITHM

The self-normalized magnitude spectrum obtained from the windowed s3(n) is generally a very smooth function with peaks only at the harmonics of the pitch. The peaks are of unequal size, the larger ones showing up at the resonance of

the first formant. In the case of the phoneme /i/ for example, a vowel with an extremely low F1 frequency, the first harmonic is generally very large compared to all of the others. The back vowel /a/ on the other hand, generally has a more graceful bulge in the high end of the spectrum, with the largest peak near 800 Hz or so (Figure 3).

The variability in size of peaks would not be a problem if there were never any spurious peaks. Unfortunately, such is not the case, for the speech waveform never behaves in any guaranteed fashion. A common problem is the presence of subharmonic peaks in the spectrum half way between the true harmonics, possibly caused by irregularities in the laryngeal excitation. These are nearly always smaller than their neighbors, but they may very well not be smaller than other true harmonics not at the formant resonance. Thus a simple measure of distance between peaks above a fixed threshold may yield a better score for a pitch choice in Hz of half the true value. A further serious problem with telephone speech is that the carrier cosine often contains 60 Hz interference which shows up as 60 Hz modulation of the speech waveform. The consequence of such interference is spurious peaks on either side of a large peak, 60 Hz away. These are often larger than true harmonics not at the formant resonance (Figure 4).

Another fact which increases the difficulty of pitch detection is the wide variability in the number of peaks to expect to find. For a high pitched female voice, there are often only two peaks which should even be considered, and the pitch is the distance between them. For an 80 Hz male voice, on the other hand, one expects to find at least 10 peaks at the harmonics of the pitch. An algorithm has to recognize the fact that there may be only two valid peaks, yet most of the time it should consider far more than two peaks in making a decision.

The algorithm described here uses an iterative technique which begins by considering only the two largest peaks. It then adds each peak in turn, from largest to smallest, and after the addition of each new peak determines a new list of potential pitches as the distance between adjacent peaks under consideration. Such a technique results in a built-in weighting mechanism, whereby the largest peak is included in every iteration, but the smallest only in the last. The final decision algorithm determines the pitch from a list which includes all of the estimates from each iteration.

The first step in extracting the pitch is to find all peaks in the spectrum and to eliminate from consideration those which are judged to be spurious. For each peak, an

amplitude and a frequency location are determined. The location is defined simply as the frequency at which the actual peak occurs. The amplitude is defined not as the magnitude of the sample at the peak, but rather as the "area under the hump". That is to say, the amplitude of a given peak is the non-normalized sum of the amplitudes of all of the samples from the previous valley to the following valley. In the event that the sum overflows 16 bits, it is clamped at +1. This choice of definition was found to effect a better separation between true peaks and spurious peaks than would a simple amplitude at the peak.

Peaks are eliminated from consideration if they are too small and/or too close to a neighboring peak. Specifically, a peak is removed if its location is within 6 samples (40 Hz) of a larger neighboring peak. A peak which is more than 6 but fewer than 10 samples away from its nearest neighbor is removed if its amplitude is less than 1/2 the amplitude of the near neighbor.

The peaks that remain after the elimination step are given a rank order according to size. At the first iteration, a single pitch estimate is entered into a table of potential pitch estimates, defined as the distance between the two largest peaks. At the second iteration, the third largest peak is added to the list of peaks under

consideration and two new pitch stimates are added to the table, defined as the distance between adjacent peaks, among the three under consideration. At each subsequent, ith, iteration, the largest peak among those remaining is added to the list of candidate peaks, and i new pitch estimates are added to the growing list of estimates, defined, again, as distance between adjacent peaks (Figure 5).

Pitch estimates are always added to the table in order, with the smallest at the beginning of the table. After each iteration, a score is computed for the maximum number of consecutive "equal" pitch estimates in the table. [Equal is defined as within 14 Hz of the succeeding entry in the table.]

As soon as there are at least 6 "equal" estimates, the average value for the "equal" entries is defined as the pitch (in Hz). If there are fewer than 6 "equal" estimates, then the algorithm continues with the next iteration until the size of the next available left-over peak is less than 1/10 the size of the largest peak, or until a maximum of 7 peaks have been exhausted. If either of these conditions is met, the algorithm exits in spite of an inadequate score, and chooses as the pitch value the average of the longest string of "equal" estimates. In the case of a tie between two strings, the one with the larger pitch estimate is

arbitrarily defined to be the pitch.

This harmonic detection algorithm is run twice per 20 msec frame on spectra of data spaced by 10 msec intervals. The output is thus an oversampled, unsmoothed pitch contour, and the final step in the processing is to make the buzz-hiss decision and decide a single pitch value for each frame. For this purpose, the pitch contour is passed through a three point followed by a five point median smoothing filter<8>(Figure 6).

The buzz-hiss decision is made almost exclusively on the basis of the smoothness of the pitch contour. Since the only true feature distinguishing voiced from unvoiced speech is the presence of pitch pulses, and since the linguistic and acoustic constraints on the pitch make it highly unlikely for a true pitch value to change dramatically in the course of a ten millisecond interval, one can expect that in voiced regions the pitch will change little from sample to sample. In unvoiced regions, on the other hand, there is little reason to expect the algorithm to arrive at anything other than random values for the pitch choice. The only other feature used by the buzz-hiss decision is an extremely conservative silence threshold on the doubly downsampled waveform, s3(n).

Thus the buzz-hiss decision operates as follows. If the energy in s3(n) is less than the silence threshold, consider the frame hiss and set the pitch equal to 0. If none of the three input samples to the three point median smoothing filter are "equal" (where "equal" is here defined as within 33 Hz of each other) consider the output of the median smoother to be 0 (hiss) Finally, if no more than 2 of the 5 ordered input samples to the 5 point median smoother are "equal" (this time within 20 Hz of each other), consider the output of the 5 point smoother to be 0 (hiss). (Figure 7).

This algorithm works surprisingly well for determining buzz-hiss. It depends upon a 10 msec rather than 20 msec update of the pitch. Typical buzz-hiss indicators such as zero crossing density, R1/R0, and high-low energy ratios were avoided on purpose, because these are likely to be degraded as a consequence of filters, distortions, and noise to which the input speech may have been subjected.

IV. LDVT IMPLEMENTATION

The algorithm as described above was incorporated into a real time linear prediction vocoder implemented on the

Lincoln Digital Voice Terminal. The LDVT is a 55 nanosecond instruction cycle microcomputer, with a standard instruction set, designed and built at Lincoln Laboratory. Memory size is a limiting factor with the machine, for it has only 2000 octal program and 1000 octal data memory locations. There is, however, a rapid access outboard memory containing 4000 octal locations from which both programs and data can be retrieved.

The preemphasized analog waveform was filtered and sampled at 132 usec intervals, and a non-overlapping buffer of 153 samples was accumulated for each 20 msec frame. These 153 samples were used as input both to the autocorrelator and to the first FIR filter, FIR1 (refer to Figure 1). The 51 output samples of FIR1 were complex multiplied by $e^{-j\pi n/4}$ and processed again through the FIR filter, using certain tricks to handle the complex input data, to yield 17 new samples of s3(n). The 128 point FFT was computed twice per frame by moving along alternately by 9, then 8, samples of s3(n). The computation of the magnitude spectrum from the 32 most recent samples of s3(n), padded out with zeros to 128, completed the preprocessing.

For the postprocessing, a table of peak locations and corresponding table of amplitudes was determined and arranged in descending order with respect to peak size.

Following this step, the first two location entries were reordered and the difference between the two locations was entered as the first pitch estimate. Then the third entry in the location table was inserted in order and two new pitch estimates, defined as difference between adjacent entries, were added, also in order, to the growing pitch estimate table. Now the 3 ordered entries in the estimate table could be scored for "equality" of adjacent elements, and an iteration is completed. At each ith iteration the ith location is inserted in order and i new pitch estimates are added in order to the estimate table. Processing is complete either when a peak of insufficient amplitude is encountered, or a score of greater than 7 adjacent "equal" estimates is obtained. At this point the mean value of the "equal" set is defined as the (unsmoothed) pitch.

An appreciation of the complexity of the algorithm can be gained from some numbers associated with the LDVT implementation. The total number of memory locations required for the entire pitch algorithm was 1425 decimal, divided about fifty fifty between instructions and data. The amount of time consumed for the preprocessing (FIR filters and computation of magnitude spectrum) was 2.66 msec per 10 msec frame, or a little over a quarter of the time available. The time required for the postprocessing, or decision algorithm, was extremely variable, and therefore

difficult to determine, but a rough calculation indicates that it was insignificant compared to preprocessing time. For purposes of comparison, the total time requirement was roughly twice the amount required by the LDVT implementation of the Gold-Rabiner time domain pitch detector.

V. RESULTS

The harmonic pitch detector, incorporated into a real time 4000 bits/sec LPC vocoder, was evaluated subjectively by means of an AB comparison with the Gold-Rabiner time domain detector<2>, incorporated into an otherwise identical vocoder. A system was developed on the UNIVAC 1219 facility whereby the two vocoders could be swapped into the LDVT essentially instantaneously, while speech subjected to various distortions and corruptions was continuously being played. The listener could thus, because of the instantaneous juxtaposition, readily compare the quality of the speech produced using the frequency domain and the time domain pitch detector.

Input speech subjected to typical telephone channel degradations was generated by means of a second LDVT containing a real time digital telephone channel simulator <11> (Figure 8) The parameters of the simulator were

controlled at the console and thus the user could conveniently test the performance of the two pitch detectors, with increasing amounts of various corruptions. For example, if one wished to investigate the sensitivity of the two pitch detectors to Gaussian noise, one could set all parameters of the telephone simulator to zero except the Gaussian noise. The noise amplitude could then be slowly increased while the two pitch detectors were alternately loaded into the other LDVT.

Using this experimental setup, we were able to examine the relative sensitivity of the two pitch detectors to the various distortions in the telephone lines. The major source of breakdown in the Gold-Rabiner pitch detector is the telephone band pass filter, which removes information below 300 Hz, attenuates the amplitude up to as much as 1000 Hz, and changes the phase relationship. Subjective listening tests show a substantial improvement in quality when the harmonic pitch detector is substituted for the time domain detector, under conditions when only the telephone filter is present in the simulation. Figure 9 shows an example where the periodicity is not evident in the waveform, but is well indicated in the spectrum, when the speech is processed through a typical telephone filter.

Other audible degradations in typical telephone lines are Gaussian noise (thermal noise and shot noise) and phase jitter. The latter is a low frequency modulation of the waveform as a consequence of (usually 60 Hz) interference in the generation of the carrier cosine. In some European lines the 50 Hz jitter can be as high as 35 degrees peak to peak amplitude, causing a peculiar granular quality and an echo effect in the speech.

Both detectors were sensitive, as might be anticipated, to Gaussian noise, although the breakdown as a consequence of Gaussian noise was not as great as might be expected. The Gold-Rabiner detector was far more sensitive to the telephone filter alone than to Gaussian noise alone, set at the level typically encountered in telephone lines (-40dbmc). The two detectors were judged to be about equally sensitive to Gaussian noise.

Phase jitter contributes an additional degradation to the time domain detector, particularly at the levels encountered in European lines. Included in the harmonic pitch detector decision algorithm is a step to suppress peaks too close to neighbors and of insufficient amplitude, which makes the detector less sensitive to phase jitter than the time domain detector. At typical American line settings, phase jitter presents little problem to either

detector.

The remaining parameters in the simulator, with the possible exception of harmonic distortion, seem to have little effect on pitch extraction, at the levels commonly found in the telephone system.

The two pitch detectors were also evaluated on certain other types of degraded speech. Specifically, speech in the presence of a) helicopter noise, b) noise in a large jet airplane, and c) 60 Hz hum, was processed through both vocoders, and the quality was compared. Helicopter noise was found to be concentrated in frequencies above 1000 Hz, and therefore caused only minor degradations in both detectors. Jet noise includes a large component in the low frequency region (below 300 Hz) and therefore interferes rather severely with the time domain pitch extraction algorithm. The same is true, obviously, for 60 Hz hum, whose strongest component is at 60 Hz, but which contains weaker harmonics at higher frequencies.

For both the 60 Hz hum and the jet engine noise, the harmonic pitch detector performed substantially better than the time domain detector. Even at levels of hum in which the time domain detector completely broke down, chosing 60 Hz as the pitch, the harmonic detector came through with clear speech. Figure 9b shows an example where the pitch

information is obscure in the waveform but evident in the spectrum, when the speech is corrupted with large airplane jet noise.

For one specific kind of distortion in transmission channels, the harmonic detector can actually correct the distortion and improve the quality of the original speech. This is for the situation in which there happens to be a very large frequency offset between the transmitter and receiver carrier in a single side band transmission system. In such a case, both positive and negative frequency are shifted in towards the origin by an amount equal to the offset, such that the original pitch harmonics are no longer harmonics. The subjective result is that the perceived pitch is wrong, and a small amplitude background hum is heard at the correct pitch. The harmonic pitch detector, since it does not depend upon the fundamental but only upon spacing between harmonics, can restore the original speaker's pitch in the synthesized speech, and remove the background hum. The formant frequencies are of course still shifted, but the formant shift is a second order effect, perceptually.

VI. SUMMARY

A frequency domain pitch detector was described which extracts the pitch information from the spacing between harmonics in a selected portion of the spectrum. The algorithm was developed with the following design criteria: a) it should perform well when the input is telephone quality speech, and b) it should be implementable in real time on a standard fast microprocessor. The algorithm so developed not only met its design goals, but also was found to obtain accurate pitch in the presence of a wider range of noise and distortions, including 60 Hz hum, jet engine noise, and large frequency offsets.

Looking to the future, the algorithm has the further advantage that it is potentialy easily implementable in CCD hardware, as the entire preprocessing consists of FIR filters and FFT's, both suitable for CCD implementation.

ACKNOWLEDGEMENT

The author would like to thank Dr. Ben Gold for sustained interest and valuable insights offered during many fruitful discussions.

REFERENCES

<1> P.E. Blankenship, et al., "The Lincoln Digital Voice Terminal System," Technical Note 1975-53, Lincoln Laboratory, M.I.T. (25 August 1975), DDC AD-A017569/5.

<2> B. Gold and L.R. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain," J. Acoust. Soc. Am. 46, 442-448 (1969).

<3> C.M. Harris and M.R. Weiss, "Pitch Extraction by Computer Processing of High Resolution Fourier-Analysis Data," J. Acoust. Soc. Am. 35, 339-343 (1963).

<4> J.D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation," IEEE Trans. Audio Electroacoust. AU-20, 367-377 (1972).

<5> R.L. Miller, "Performance Characteristics of an Experimental Harmonic Identification Pitch Extraction (HIPEX) System," J. Acoust. Soc. Am. 47, 1593-1601 (1970).

<6> J.A. Moorer, "The Optimum Comb Method of Pitch Period Analysis of Continuous Digitized Speech," IEEE Trans. Acoust., Speech, and Signal Processing, ASSP-22, 330-338 (1974).

<7> A.M. Noll, "Cepstrum Pitch Determination," J. Acoust. Soc. Am. 41, 293-309 (1967).

<8> L.R. Rabiner, M.R. Sambur, and C.E. Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing," IEEE Trans. Acoust., Speech, and Signal Processing, ASSP-23, 552-557 (1975).

<9> M.J. Ross, H.L. Shaffer, A. Cohen, R. Freudberg, H. Manley, "Average Magnitude Difference Function Pitch Extractor," IEEE Trans. Acoust., Speech, and Signal Processing, ASSP-22, 353-362 (1974).

<10> M.R. Schroeder, "Period Histogram and Product Spectrum: New Methods for Fundamental Frequency Measurement," J. Acoust. Soc. Am. 43 (1968).

<11> S. Seneff, "A Real-Time Digital Telephone Simulation on the Lincoln Digital Voice Terminal," Technical Note 1975-65, Lincoln Laboratory, M.I.T. (30 December 1975), DDC AD-A021409/8.

<12> M.M. Sondhi, "New Methods of Pitch Extraction," IEEE Trans. Audio Electroacoust. AU-16, 262-266 (1968).

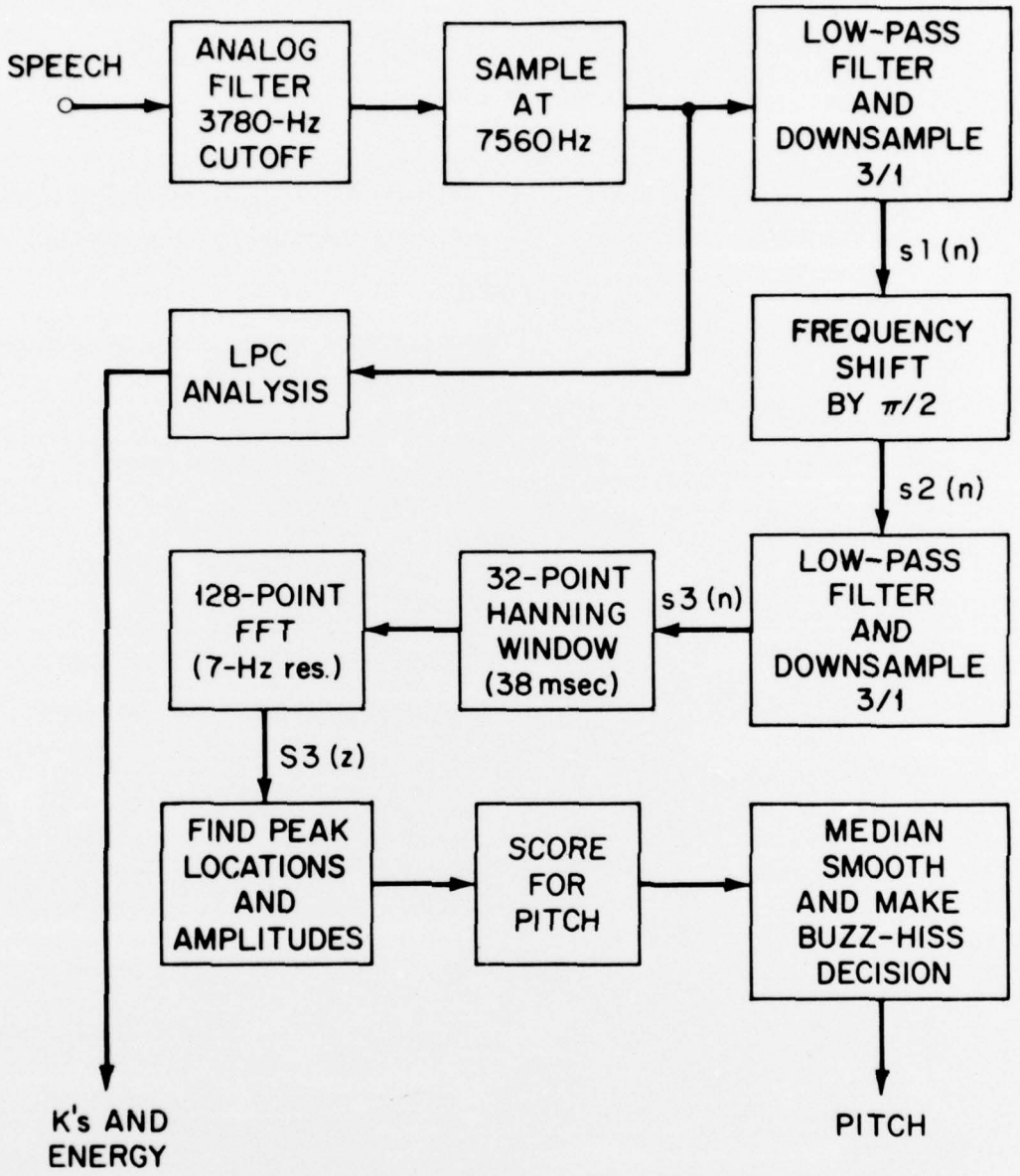
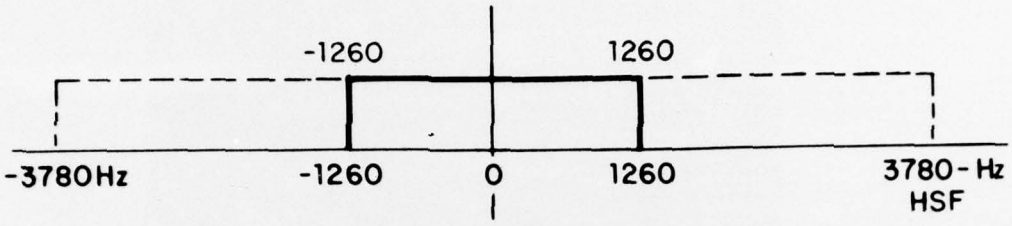
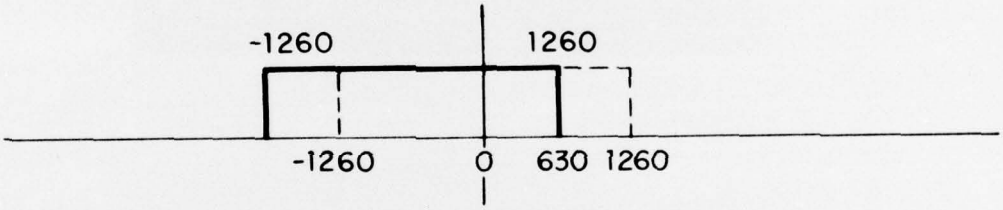


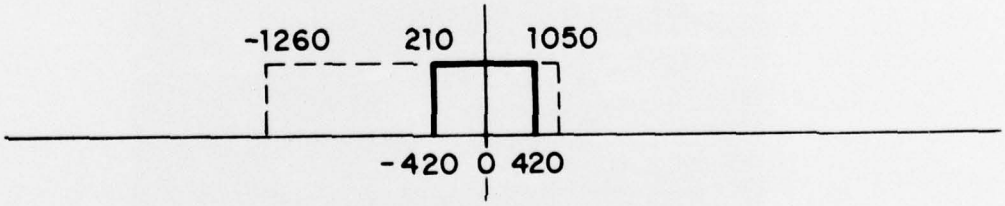
Fig. 1. Block diagram of harmonic pitch detector.



(1) LP FILTER AND DOWNSAMPLE 3/1

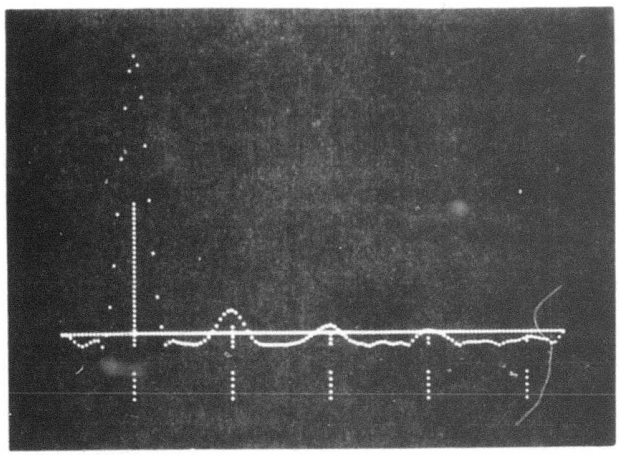


(2) SHIFT SPECTRUM: MUL EACH s(n) BY $e^{j\pi n/2}$

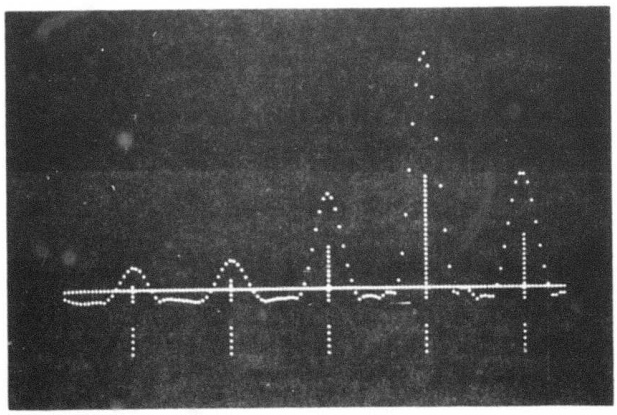


(3) LP FILTER AND DOWNSAMPLE 3/1

Fig. 2. Preprocessing of speech waveform to obtain downsampled signal with desired spectral information.



/i/



/a/

Fig. 3. Typical first format region spectra for the vowels /i/ (above) and /a/.

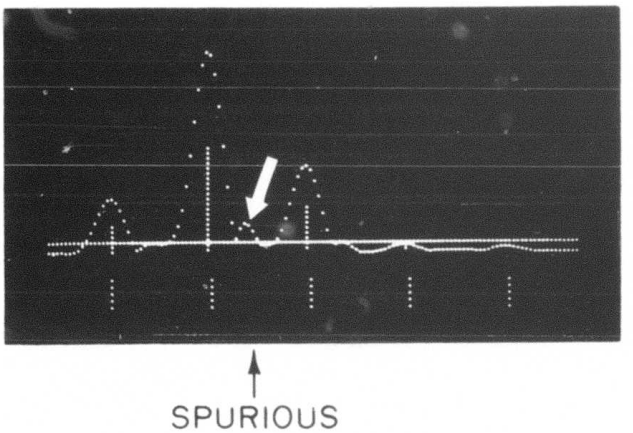


Fig. 4. Introduction of spurious peak in spectrum as a consequence of 60 Hz phase jitter.

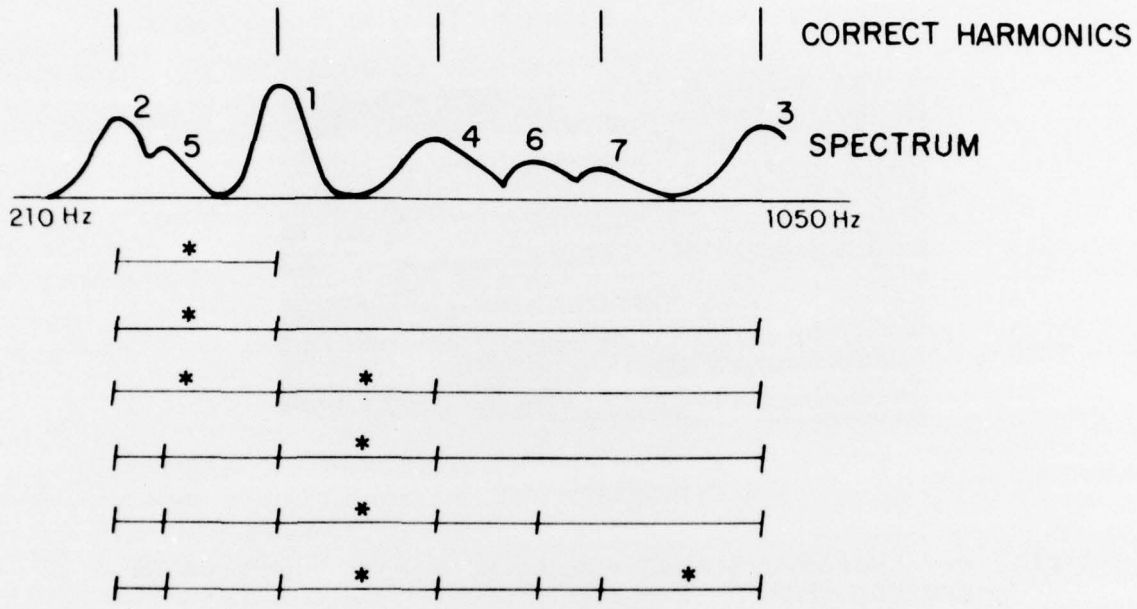


Fig. 5. Illustration of iterative scoring algorithm under artificially adverse conditions.

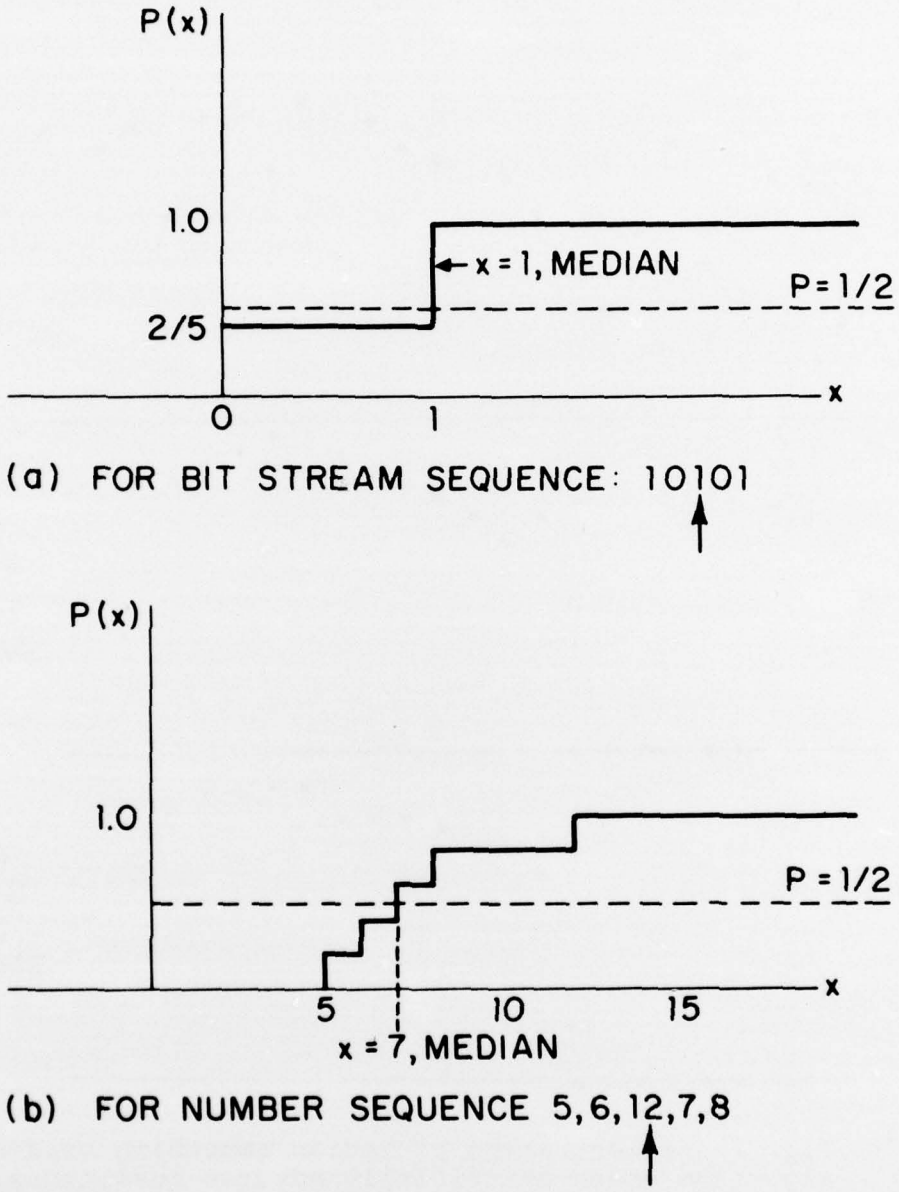


Fig. 6. Median smoothing filter, a) for bit stream and b) for function.

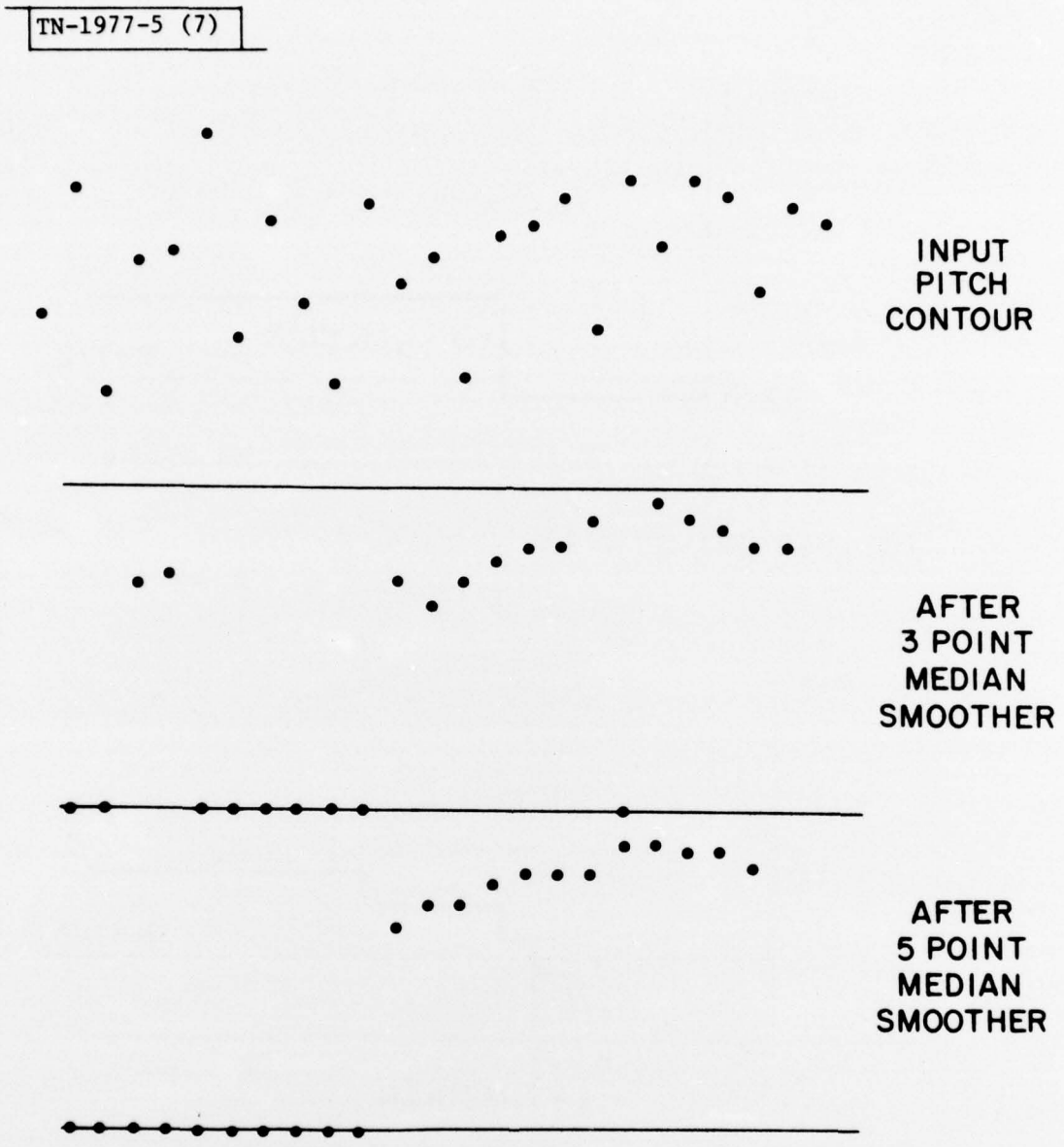


Fig. 7. Illustration of median smoothing buzz-hiss algorithm under artificially adverse conditions.

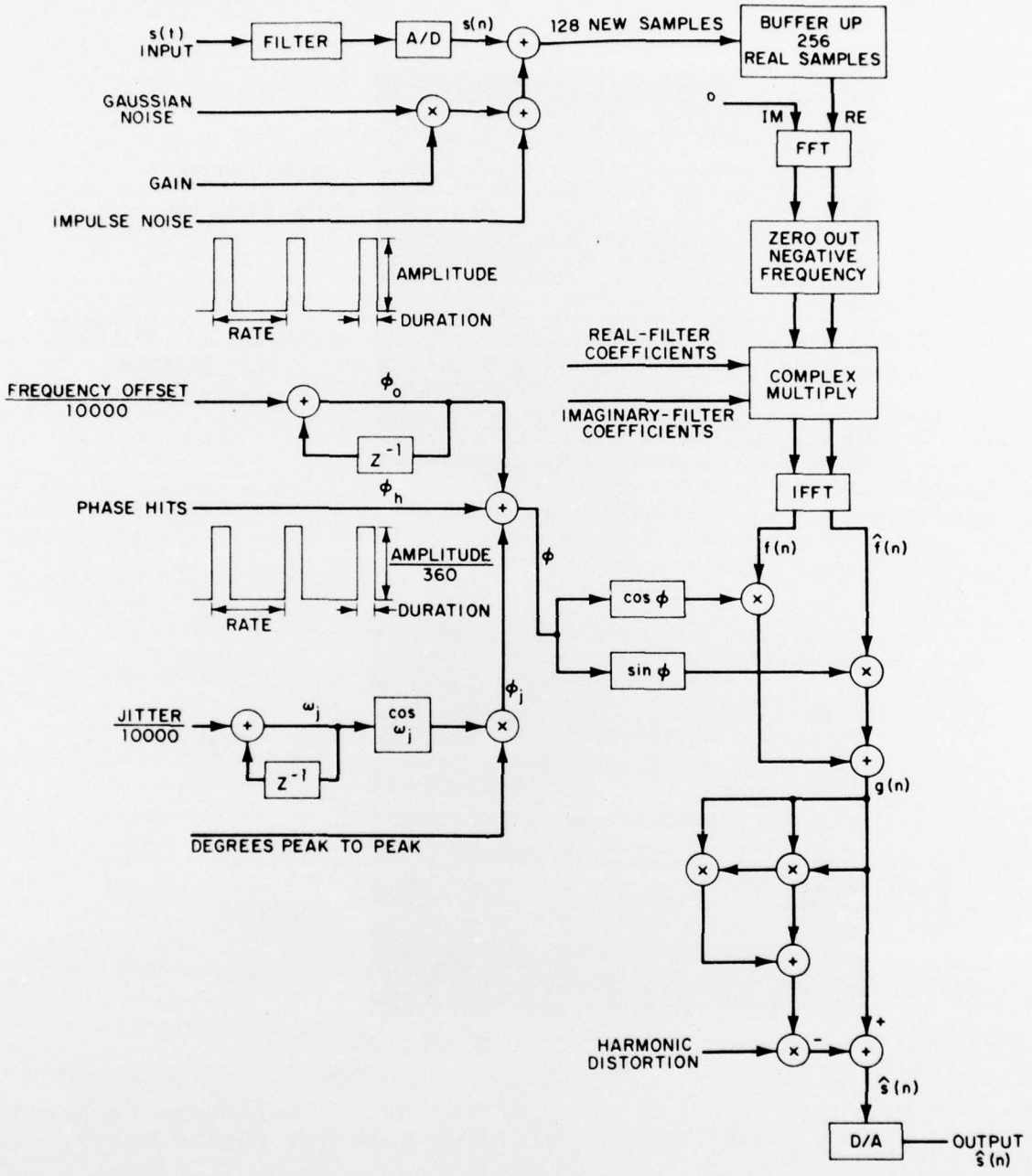
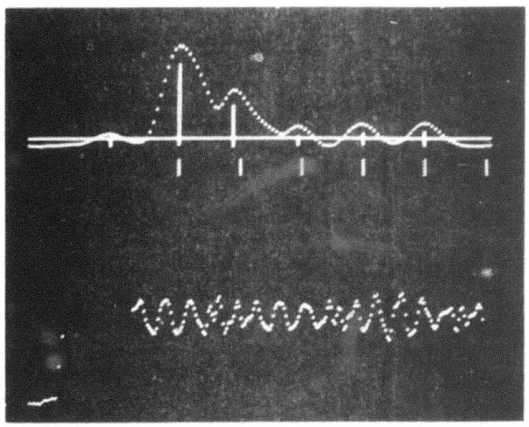
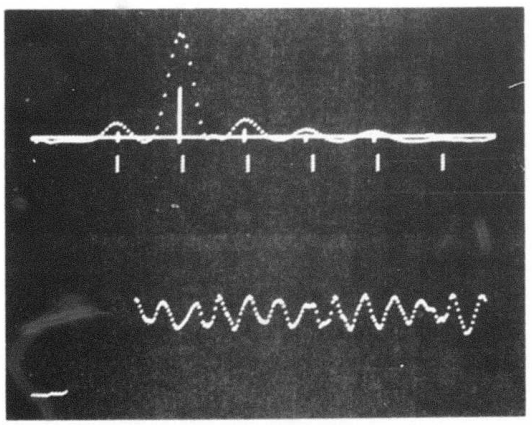


Fig. 8. Block diagram of telephone channel simulator used to test harmonic pitch detector performance.



(a)



(b)

Fig. 9. Waveform and spectrum of, a) telephone filtered speech and b) speech corrupted with jet engine noise.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: ESD-TR-77-35
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)*: Real Time Harmonic Pitch Detector
5. TYPE OF REPORT & PERIOD COVERED: Technical Note
6. PERFORMING ORG. REPORT NUMBER: Technical Note 1977-5
7. AUTHOR(s): Stephanie Seneff
8. CONTRACT OR GRANT NUMBER(s): F19628-76-C-0002, ARPA Order-2006
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Lincoln Laboratory, M.I.T., P.O. Box 73, Lexington, MA 02173
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: ARPA Order 2006, Program Element No. 62706E, Project No. 7P10
11. CONTROLLING OFFICE NAME AND ADDRESS: Defense Advanced Research Projects Agency, 1400 Wilson Boulevard, Arlington, VA 22209
12. REPORT DATE: 26 January 1977
13. NUMBER OF PAGES: 40
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*: Electronic Systems Division, Hanscom AFB, Bedford, MA 01731
15. SECURITY CLASS. *(of this report)*: Unclassified
15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

TN-1977-5

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

harmonic pitch detection
Lincoln Digital Voice Terminal (LDVT)
vocoders
speech waveforms

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

A real time harmonic pitch detection algorithm has been developed on the Lincoln Digital Voice Terminal (LDVT). The algorithm was designed to be fast and to perform well when the input speech is degraded (i.e., telephone quality) or corrupted with acoustically coupled noise. The algorithm determines the fundamental frequency from the spacing between harmonics in a selected portion of the spectrum. The algorithm was incorporated into a real time linear prediction vocoder and compared favorably in informal listening tests with the Gold-Rabiner time domain detector under a variety of adverse conditions.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

204650